Replication Package for
"Emotion and Reason in Political Language"
Gloria Gennaro, Elliott Ash

This document describes the files needed to reproduce the results reported in:

> Gennaro, Gloria, and Ash, Elliott. 2021. "Emotion and Reason in Political Language". *The Economic Journal*.

In particular, it includes all the datasets, models and codes to reproduce the figures and tables in the paper and in the appendix. For reference, we also provide all the code that was used to produce the main dataset, including all the variables that derive from text data.

Folder Description

The replication package includes four folders: "scripts", "data", "models" and "results".

The "scripts" folder includes the following material:

- 0_data_creation: the folders include all Python scripts that have been used to produce the main dataset from raw files. These are numbered sequentially, and should be run in this order. These files should be considered as a reference for the reader who wants to implement a similar emotionality measure in a different context. Running those scripts requires access to the Congressional Records and to the LWIC word lists.
- 1_main_paper:
  - figures: the folder includes individual Python scripts to produce all paper figures. Numbers correspond to the figure numbering. Expected running times are:
    - *fig1.py:* 9.84s
    - *fig2.py:* 51.31s
    - *fig3.py:* 68.81s
    - *fig4.py:* 76.36s
    - *fig5.py*: 57.91s
  - tables: the folder includes individual R and Stata scripts to produce all paper tables. Numbers correspond to the table numbering (*tab1.R – tab2.do*). The file *0_data_cleaning.do* needs to be run once before all table creation files – This step can be skipped by uploading directly the "dataset_tables.dta" in the data directory. Running times are:
    - *0_data_cleaning.do*: 6m48s
    - *tab1.R*: 0.83s
    - *tab2.do*: 10m16s
- 2_appendix:
  - figures: the folder include individual Python and Stata scripts to produce all appendix figures. Numbers correspond to the figure numbering (*figA1a.py – figA21.do*). Running times are:
    - figA1.py: 11.93s
    - figA2.py: 52.80s
    - figA3.py: 65.98s
    - figA4.py: 96.44s

        - figA5.py: 62.80s
        - figA6.py: 60.17s
        - figA7.py: 75.37s
        - figA8-9_prep.do: 5m53s
        - figA8.py: 1.62s
        - figA9.py: 1.16s
        - figA10.py: 67.83s
        - figA11.py: 62.57s
        - figA12.py: 62.73s
        - figA13.py: 76.48s
        - figA14.py: 62.18s
        - figA15.py: 87.40s
        - figA16.py: 101.84s
        - figA17.py: 96.73s
        - figA18.py: 99.14s
        - figA19.py: 81.00s
        - figA20.py: 80.86s
        - figA21.do: 1m42s
    - tables: the folder includes individual Python, R and Stata scripts to produce all appendix tables. Numbers correspond to the table numbering (*tabA1.py – tabA10.do*). Running times are:
        - tabA1.py: 206.66s – runs on synthetic data
        - tabA2.py: 105.31s – runs on synthetic data
        - tabA3.py: 105.22s – runs on synthetic data
        - tabA4.do: 1m10s
        - tabA5.py: 158.42s – runs on synthetic data
        - tabA6.R: 0.83s
        - tabA7.py: 89.80s
        - tabA8.do: 20m01s
        - tabA9.do: 10m29s
        - tabA10.do: 12m53s

The "data" folder includes the following material:

1. 1_main_datasets: includes the two datasets necessary for reproducing all tables and figures in the paper and appendix, and a variable list:
    a. *main_dataset.csv* includes all variables used in the study, plus additional contextual variables.
    b. *dataset_tables.dta* is the output of the script *0_data_cleaning.do.* It includes all the variables that are necessary to reproduce the results in the regression tables.
    c. *variable_list.xls* includes a variable list with description.
2. 2_validation_results: includes the result of the human validation exercise. Since this was performed in two batches, we report them separately (*validation_round1.csv, validation_round2.csv)* as well as jointly (*validation_consolidated.csv*) to facilitate the replication of the paper tables
3. 3_auxiliary_data:
    a. affect_centroid.pkl: embedding vector for the affect pole
    b. cog_centroid.pkl: embedding vector for the cognition pole
    c. dictionary_affect.pkl: final affect word list
    d. dictionary_cognition.pkl: final cognition word list
    e. emotionality_score_res_demo_topics.csv: intermediary output of figA8-9_prep.do

f. *emotionality_score_res_demo.csv*: intermediary output of figA8-9_prep.do
g. *procedural_words.pkl:* full list of procedural words dropped from the analysis
h. *stopwords.pkl:* full list of stopwords words dropped from the analysis
i. *topics_numbers.pkl:* list of topics indexed with a reference number, topic level and keywords
j. *word_counts.pkl:* absolute word frequencies for each vocabulary word in the corpus
k. *word_freqs.pkl:* relative word frequencies for each vocabulary word in the corpus
l. *synth_speeches_tabA1.pkl:* sample of speeches to reproduce tables A1 and A5
m. *synth_speeches_tabA2.pkl:* sample of speeches to reproduce table A2
n. *synth_speeches_tabA3.pkl:* sample of speeches to reproduce table A3

The "models" folder includes the trained word embedding model:

- *w2v-vectors_8_300.pkl*
- *w2v-vectors_8_300.pkl.trainables.syn1neg.npy*
- *w2v-vectors_8_300.pkl.wv.vectors.npy*

The "results" folder includes the following material:

- main_paper: all figures and tables produced by running the scripts in "*scripts/1_main_paper"*
- appendix*:* all figures and tables produced by running the scripts in "*scripts/2_appendix"*

Software and module versions

- macOS Catalina, version 10.15.7

- Stata/SE 16.1 for Mac (64-bit Intel)
  - Packages: binscatter, cmogram, reghdfe

- R 4.0.5 (2021-03-31)
  - R packages: stringr (1.4.0), stringi (1.6.2), plyr (1.8.6), tidyr (1.1.3), data.table (1.14.0), dplyr (1.0.6)

- Python 3.8.5
  - Python packages: matplotlib (3.3.2), nltk (3.5), wordcloud (1.8.1), joblib (1.0.0), genism (3.8.3), pandas (1.2.1), numpy (1.19.5), scipy (1.5.2)
  - In case of issues with genism, use version 3.4.0

Data citations

Ash, Elliott, Massimo Morelli, and Richard Van Weelden. "Elections and divisiveness: Theory and evidence." *The Journal of Politics* 79.4 (2017): 1268-1285.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science*, 331.6014 (2011): 176-182

Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2021). *Voteview: Congressional Roll-Call Votes Database.*

CQ Congress Collection Database, 79th to 113th Congress.

United States Congressional Record, 35th to 113th Congress.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. "Linguistic inquiry and word count: LIWC 2001." *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001): 2001.